

Prostate Cancer Risk Prediction and Online Calculation Based on Machine Learning Algorithm

Chun Wang¹, Qinxue Chang¹, Xiaomeng Wang¹, Keyun Wang¹, He Wang², Zhuang Cui^{1*}, Changping Li^{1*}

¹Department of Health Statistics, School of Public Health, Tianjin Medical University, Tianjin 300070, China.

²Department of Medical Imaging, Peking University First Hospital, Beijing 100034, China.

ABSTRACT

Objective To build a prostate cancer (PCa) risk prediction model based on common clinical indicators to provide a theoretical basis for the diagnosis and treatment of PCa and to evaluate the value of artificial intelligence (AI) technology under healthcare data platforms.

Methods After preprocessing of the data from Population Health Data Archive, smoothly clipped absolute deviation (SCAD) was used to select features. Random forest (RF), support vector machine (SVM), back propagation neural network (BP), and convolutional neural network (CNN) were used to predict the risk of PCa, among which BP and CNN were used on the enhanced data by SMOTE. The performances of models were compared using area under the curve (AUC) of the receiving operating characteristic curve. After the optimal model was selected, we used the Shiny to develop an online calculator for PCa risk prediction based on predictive indicators.

Results Inorganic phosphorus, triglycerides, and calcium were closely related to PCa in addition to the volume of fragmented tissue and free prostate-specific antigen (PSA). Among the four models, RF had the best performance in predicting PCa (accuracy: 96.80%; AUC: 0.975, 95% CI: 0.964-0.986). Followed by BP (accuracy: 85.36%; AUC: 0.892, 95% CI: 0.849-0.934) and SVM (accuracy: 82.67%; AUC: 0.824, 95% CI: 0.805-0.844). CNN performed worse (accuracy: 72.37%; AUC: 0.724, 95% CI: 0.670-0.779). An online platform for PCa risk prediction was developed based on the RF model and the predictive indicators.

Conclusions This study revealed the application value of traditional machine learning and deep learning models in disease risk prediction under healthcare data platform, proposed new ideas for PCa risk prediction in patients suspected for PCa and had undergone core needle biopsy. Besides, the online calculation may enhance the practicability of AI prediction technology and facilitate medical diagnosis.

Key words: prostate cancer; random forest; support vector machine; back-propagation neural network; convolutional neural network

INTRODUCTION

Prostate cancer (PCa) is a worldwide health issue that poses a major threat to men's health^[1]. Early diag-

nosis and treatment of PCa can improve the treatment effectiveness. However, PCa and prostatic hyperplasia are difficult to distinguish based on symptoms alone.

Prostate specific antigen (PSA) has been widely employed in PCa screening, diagnosis, staging, efficacy monitoring, and prognosis evaluation as the most essential marker for the clinical diagnosis of PCa. PSA levels are affected by a variety of variables, including age and body mass index (BMI)^[2,3]. Furthermore, correct PCa diagnosis is beset by a slew of practical issues, including invasiveness and cost. Therefore, it is of great significance to use common clinical indicators to build a

Received March 21, 2022; accepted August 24, 2022; published online September 20, 2022.

*Corresponding authors: Zhuang Cui, zhuangcui417@126.com; Changping Li, changpingli417@126.com.

© The authors 2022. Published by Chinese Academy of Medical Sciences. This is an open access article attributed under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0>).

PCa prediction model that can lower medical expenses and increase early diagnosis and treatment rates.

In our current study, random forest (RF), support vector machine (SVM), back-propagation neural network (BP), and deep learning convolutional neural network (CNN) were used to build PCa risk prediction models with clinical indicators in this study, which were based on the PCa data set from the Population Health Data Archive.

METHODS

Participants

The Prostate Cancer dataset from the Population Health Data Archive (PHDA) was used in this study^[4], which included 1000 male patients, with 427 PCa cases and 573 prostatic hyperplasia cases. The flow chart is shown in **Fig. 1**.

For modeling and prediction, the following data were used: a) demographic information; b) prostate indicators; c) serum enzymes examination; d) blood biochemical indicators; and e) electrolyte indicators, the indicators for details were shown in **Table 1**.

Feature engineering

Data Cleaning

Box diagram division was used to modify or remove outliers, and multiple imputation was used to fill in the missing values using R software ("MICE" package). Regular expression was used to extract volume of core-biopsy sampled tissue and offwhite fragile tissue in biopsied samples, with the presence (or not) of prostatic hyperplasia and the volume of total prostatectomy being new variables. The average value was calculated if the volume of core-biopsy sampled tissue was measured multiple times.

New Variables

1) Estimated glomerular filtration rate (eGFR):

$$eGFR = 141 \times \min\left(\frac{Cr}{\kappa}, 1\right)^{\alpha} \times \max\left(\frac{Cr}{\kappa}, 1\right)^{-1.209} \times 0.993^{Age} \times 1.1 \quad (1)$$

the above is the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) creatinine equation^[5], where $\kappa=0.9$ for males and $\kappa=0.7$ for females, $\alpha=-0.329$ for females and $\alpha=-0.411$ for males, $\min(Cr/\kappa, 1)$ is the minimum value between Cr/κ and 1, and $\max(Cr/\kappa, 1)$ is the maximum value between Cr/κ and 1.

2) Free prostate-specific antigen ratio (rPSA):

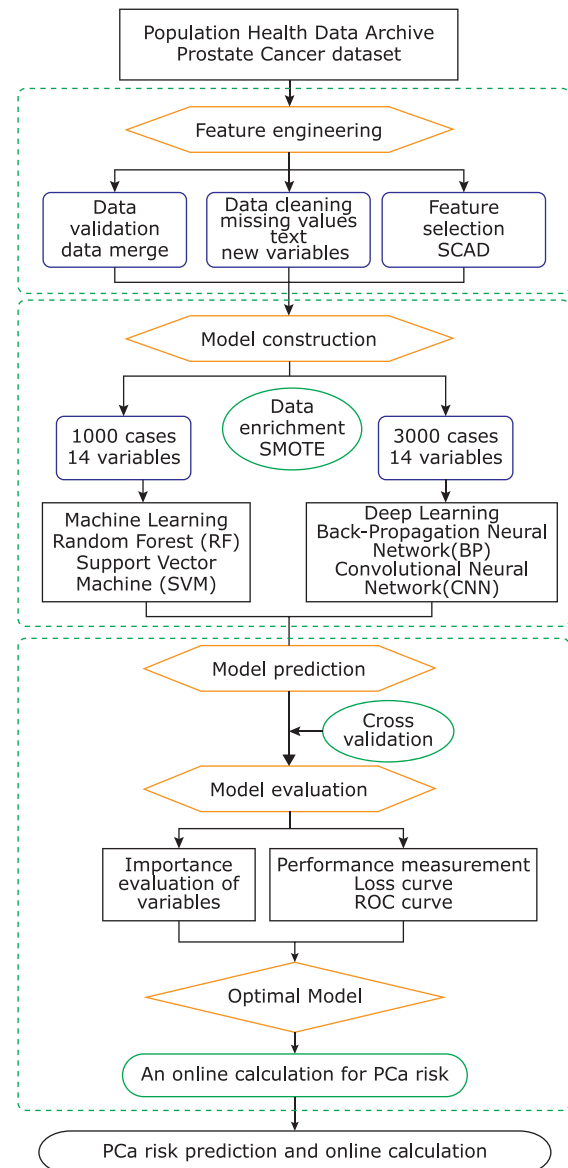


Figure 1. Research flow chart of PCa risk prediction SCAD: smoothly clipped absolute deviation; SMOTE: synthetic minority oversampling technique

$$rPSA = \frac{fPSA}{tPSA} \quad (2)$$

3) Body mass index (BMI):

$$BMI = \frac{Weight}{Height^2} \quad (3)$$

Data Enhancement

The small sample size in this study would presumably reduce the prediction power of deep learning. Therefore, the Synthetic Minority Over-sampling Technique (SMOTE) was used for data enhancement before the model construction. SMOTE can effectively counter

Table 1. The variables from the dataset and abbreviations used in modeling and prediction of prostate cancer in this study

Indicator	Abbreviation
Demographic information	
Age	-
Height	-
Weight	-
Body mass index	BMI
Prostate indicators	
Free prostate specific antigen	fPSA
Total prostate specific antigen	tPSA
Free prostate-specific antigen ratio	rPSA
Volume of core-biopsy sampled tissue	CBV
Offwhite fragile tissue in biopsied samples	OFT
Transurethral resection prostate	TURP
Presence of accompanying prostatic hyperplasia	-
Volume of prostate	PV
Serum enzymatic examination	
Alkaline phosphatase	ALP
Creatine kinase isoenzyme	CK_MB
Lactic dehydrogenase	LDH
Creatine kinase	CK
Blood biochemical indicators	
Creatinine	CR
Albuminous	ALB
Serum uric acid	UA
Triglyceride	TG
High density lipoprotein cholesterol	HDLC
Low density lipoprotein cholesterol	LDLC
Apolipoprotein A1	ApoA1
Apolipoprotein B	ApoB
Estimated glomerular filtration rate	eGFR
Electrolyte indicators	
Potassium	K
Inorganic phosphorus	iP
Sodium	Na
Inorganic calcium	iCa
Chloride	CL

class imbalance by generating new samples from the existing cases. Each minority class sample and the k-nearest neighbors have attributes in common with the synthetic sample. Therefore, without duplicating the original data, SMOTE might balance the class ratio and amplify the features of the minority class^[6].

Feature selection

Features were selected by Smoothly Clipped Absolute Deviation (SCAD). SCAD can select significant variables and estimated model coefficients. With symmetric and non-concave penalty function, SCAD can process singular matrices to create sparse solutions. Compared with ridge regression, SCAD can reduce the prediction variance of the model. Compared to Lasso,

SCAD can reduce parameter estimation bias.

Statistical analysis

Random forest

Random forest (RF) is an ensemble learning method for classification and regression problems, in which many decision trees are constructed and average predictions are produced^[7]. RF can model complex data sets without assuming that variables are regularly distributed or that other prerequisites are met, and it has a high noise and outlier tolerance, allowing it to effectively deal with complex nonlinear interactions and balance the errors of unbalanced data. In our current study, RF model was constructed in Python ("sklearn.tree" module).

Support vector machine

The basic model of support vector machine (SVM) is a classifier that defines the maximum margin in the feature space. SVM finds a hyperplane to correctly categorize training samples that are linearly separable. In real-world jobs, however, such hyperplane may not exist in the original sample space. In this case, SVM uses a kernel function to translate the samples from the input space to a higher-dimensional "Reproducing Kernel Hilbert Space", thereby separating the samples linearly separable in this feature space and effectively solving nonlinear problems^[8]. In our current study, SVM was modeled in Python ("sklearn.svm" module).

Back-propagation neural network

Back-propagation (BP) neural network, a multi-layer feedforward neural network, is characterized by signal forward propagation and error backpropagation. It can classify arbitrary complicated patterns and perform good multi-dimensional function mapping, as well as solve exclusive OR (XOR) and some other problems that a simple perceptron cannot answer. BP consists of three basic parts: input layer, hidden layer and output layer. In our current study, BP was constructed in Python ("TensorFlow" module).

Convolutional neural network

As one of the most common deep learning methods, the convolutional neural network (CNN) is a feedforward neural network with convolutional processing and a deep structure^[9]. CNN can perform supervised and unsupervised learning, as well as translation invariant categorization of input data based on its hierarchical structure. A typical CNN consists of five parts: input layer, convolutional layer, activation function layer,

pooling layer, and full connection layer. In our current study, BP and CNN were constructed in Python ("TensorFlow" module).

Dataset division and the model evaluation indicators

The test set was substituted for prediction after the prediction model was created using RF, SVM, BP, and CNN models. Deep learning models used 3000 bits of data enhanced by SMOTE for training to ensure the stability of its training and prediction, while machine learning models used original data to ensure the accuracy of their predictions. The models divided the training set and the test set according to 3:1. The evaluation of model performances included accuracy, precision, recall, and area under the receiver operating characteristic (ROC) curve (AUC). In general, a model with a substantially lower AUC value is a worse model.

RESULTS

Variable screened by SCAD

This study preliminarily appraised the strength of the association between variables and outcomes, excluded collinear variables, and used the SCAD model to filter variables in order to obtain a stable prediction model. The cross-validation error (CVE) of the model was the smallest when the 14th variable was selected, and the variable screening effect was the best at this time (**Fig. 2**). A total of 14 variables, namely BMI, ALB, CK_MB, fPSA, Ca, iP, iCa, LDH, CK, TG, LDL, ApoA1, ApoB, and CBV, were finally chosen as potential indicators for the prediction model.

Prediction of PCa outcomes in four models

RF model

The optimal parameters were set as the maximum number of features (mtry) to be 12, and the number of decision trees (ntree) to be 600. Furthermore, the minimum error model was obtained (**Fig. 3 A, B, and C**). **Fig. 3D** is the variable importance ranking diagram. A higher ranking means greater impact on the accuracy of the model. The confusion matrix of model prediction results is shown in **Table 2**.

SVM model

The SVM model was constructed with the selected variables. The best parameters were Gaussian kernel function with gamma of 0.0013, a cost of 12.5, and a total of 214 support vectors. The confusion matrix of

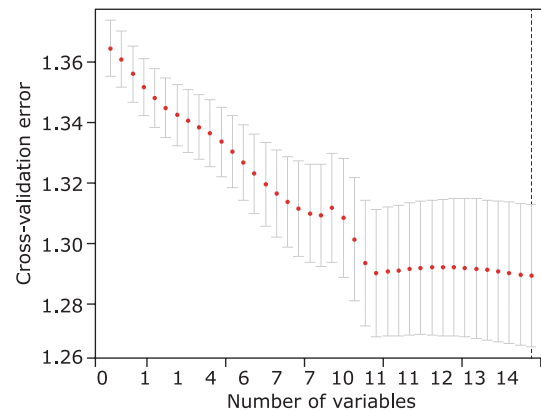


Figure 2. The screening of variables by SCAD model

The abscissa represents the number of included variables. After 39 parameter adjustments, the CVE was the smallest, which included 14 variables.

model prediction results is shown in **Table 2**.

BP model and CNN model

In general, the number of neurons in the input layer should be the same as the number of selected variables, and the number of nodes in the output layer be consistent with the study target parameters in a deep learning prediction model^[10]. In our current study, the number of nodes in the input layer was 14, the number of nodes in the hidden layer was 8, and the number of nodes in the output layer was 2. After parameter search and cross-validation, BP model was finally built by 4 hidden layers, with the activation function being "ReLU" and the optimizer being "Adam"; CNN model was finally built by 1 convolution layer, 1 pooling layer, and 3 hidden layers, with the activation function being "ReLU" and the optimizer being "Adam". The model training prediction effect is shown in **Fig. 4**. The confusion matrix of model prediction results is shown in **Table 2**.

Comparison of the prediction results of the four models

Table 3 shows that the prediction accuracies of the RF, SMV, BP, and CNN models were 96.80%, 82.67%, 85.36%, and 72.37%, and the AUC values were 0.975, 0.824, 0.892, and 0.724, respectively. RF model had the best prediction performance among these four models, followed by the BP model, the SVM model which had a comparable prediction result to the BP model, and the CNN model which performed slightly worse. The ROC curves of the four models are shown in **Fig. 5**.

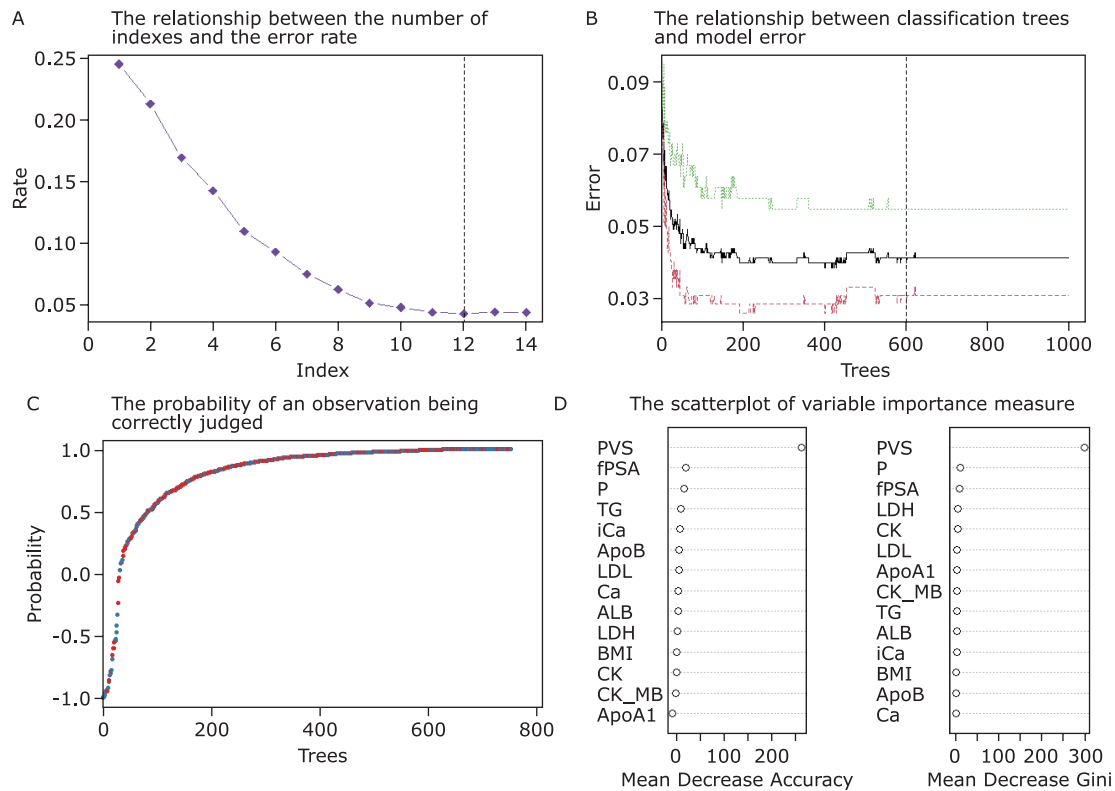


Figure 3. The visualization diagrams of the random forest (RF) model

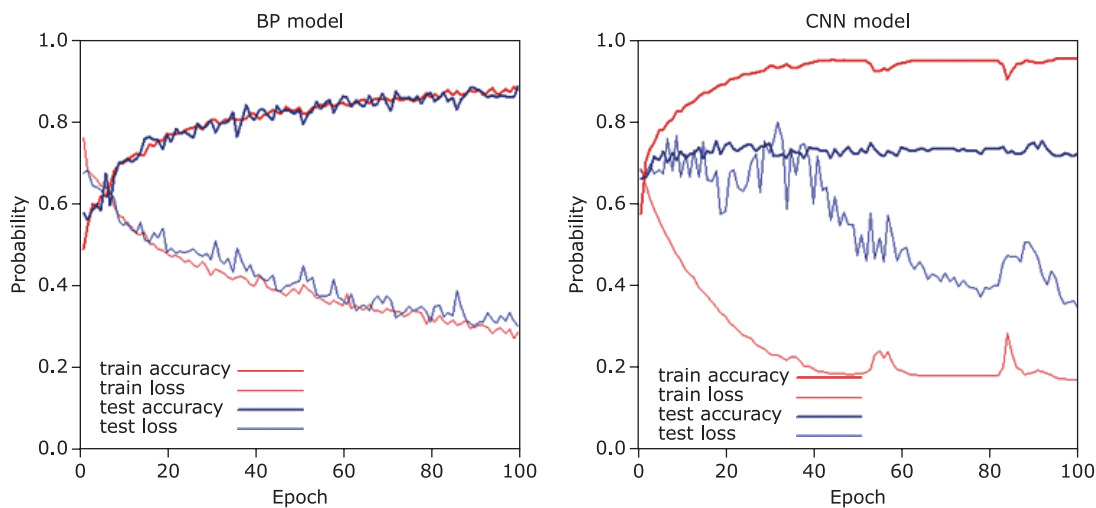


Figure 4. The training and prediction effects of the back propagation (BP) and convolutional neural network (CNN) models

PCa risk prediction online calculation based on machine learning algorithm

Based on the optimal machine learning algorithm, we selected the RF model and developed an online calculator for PCa risk prediction with the Shiny (www.shinyapps.io)^[11]. The webpage preview is shown in **Fig. 6**. The PCa risk is presented on the right side of the interface when users enter the value of each indicator and clicks "Calculate". It should be noted that, this online

calculation is only designed for preliminary PCa risk assessment and cannot replace clinical diagnosis.

DISCUSSION

Using the PCa dataset of Population Health Data Archive, we discovered 14 important variables by a literature study and SCAD, and then RF, SVM, BP, and CNN were applied to classify and predict PCa. To build models,

Table 2. The confusion matrix of prediction results of four models

Actual	Predicted		Row Total
	Non-PCa	PCa	
RF model			
Non-PCa	149	3	152
PCa	5	93	98
Column total	154	96	250
SVM model			
Non-PCa	141	26	167
PCa	26	107	133
Column total	167	133	300
BP model			
Non-PCa	264	47	311
PCa	19	270	289
Column total	283	317	600
CNN model			
Non-PCa	74	21	95
PCa	35	71	106
Column total	109	92	201

Table 3. The comparison of AUC values of four models in test set

Model	Accuracy (%)	Precision (%)	Recall (%)	AUC (95%CI)
RF	96.80	96.88	94.90	0.975 (0.964, 0.986)
SVM	82.67	80.45	80.45	0.824 (0.805, 0.844)
BP	85.36	85.17	93.43	0.892 (0.849, 0.934)
CNN	72.37	77.18	66.98	0.724 (0.670, 0.779)

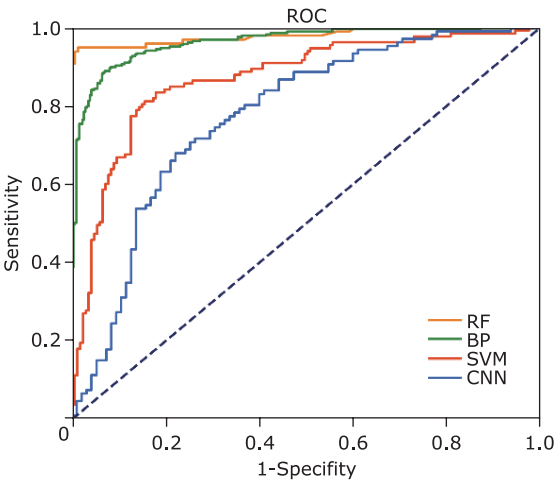


Figure 5. The comparison of ROC curves of four models

the neural networks that are better suited for big data used 3000 bits of data enhanced by SMOTE. We used a variety of evaluation indicators to measure and visualize the model's predictive ability. The best model for predicting PCa was determined, as well as the variable importance ranking of PCa. Finally, a machine learning technique was used to create an online calculation.

The best performance in the classification and prediction of PCa was achieved by RF, which differs from traditional decision trees in that it uses bagging to introduce random attribute selection during training. The purpose of BP is to minimize the accumulative error, and CNN combines multiple convolutional layers and pooling, which is essentially sub-sampling to re-

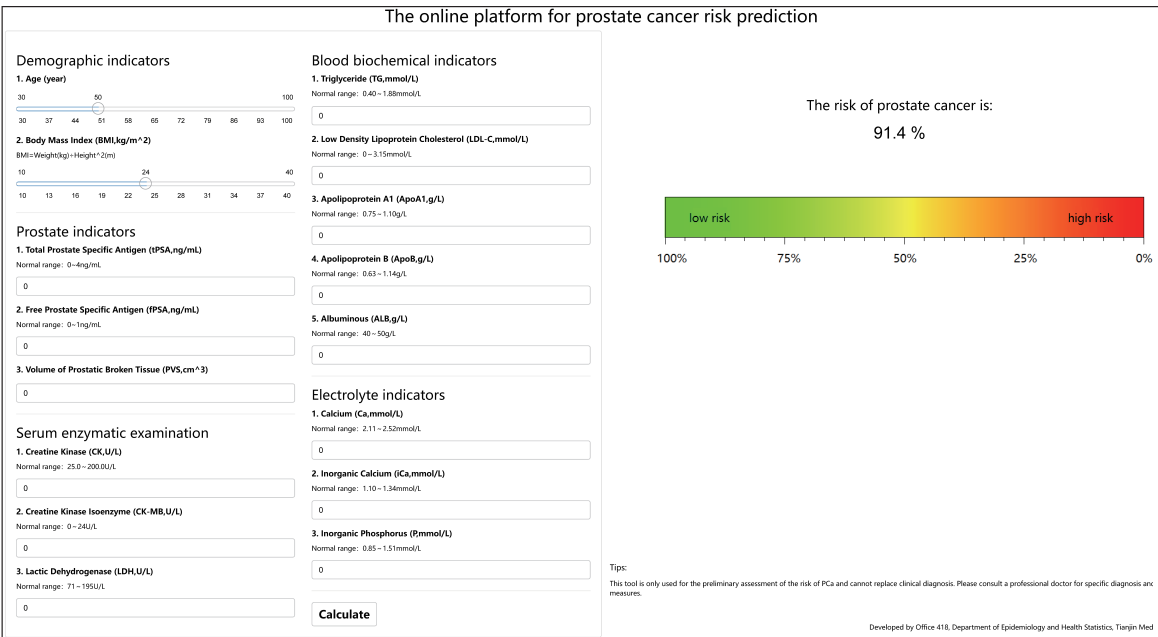


Figure 6. Interaction web page of the online platform for prostate cancer risk prediction

duce the amount of data. Even after SMOTE enhancement, the small sample data with CNN lost a lot of features in this study. SVM has a lower training performance than BP, which may be due to the fact that the current features and data are not mapped to the most appropriate feature space.

In addition, we found that CBV, fPSA, P, TG, iCa, and other clinically common biochemical markers were closely related to PCa. fPSA level rises with age and differ significantly by race^[12]. Although further research is needed to validate the link between P and PCa^[13], P has been found to be one of the most important indicators in the early diagnosis of PCa^[14]. As an organic compound, TG is one of the risk factors for PCa^[15,16]. It is possible that androgens regulate the expressions of lipid genes, resulting in the buildup of neutral lipids (TG and cholesterol) and consequently the development of PCa. iCa plays an important role in glandular secretion as it can promote the secretion of vascular endothelial growth factor and increase the transcription level of hypoxia-inducible factor related to tumor genesis and metastasis. In addition, calcium-binding protein can induce carcinogenesis^[17].

The innovation of this study is the online calculation for PCa risk based on machine learning algorithm. The visual interactive interface makes diagnosis of the diseases easier or faster in individuals who are suspected for PCa and have undergone core needle biopsy, and the online calculation can translate PCa prediction into intuitive quantitative information, giving good reference for medical diagnosis. However, our current study had several limitations. The dataset we used in developing the prediction model of PCa came from publicly available clinical data. The dataset does not include some common indicators (e.g., multiparametric prostate MRI) in diagnosis and treatment of PCa. Since the dataset is characterized by suspected PCa patients with variables from clinical settings, the target population for the prediction model was limited to patients who were suspected for PCa and had undergone core needle biopsy and associated laboratory examinations. For the predictive model, external validation is needed to determine its generalization when applying to patients from other hospitals; the indicators included in the model need to be optimized in the future to make them clinically practical for individuals at high risk of PCa.

Acknowledgement

We thank all the medical practitioners in General Hos-

pital of the People's Liberation Army for their contributions in generating and collecting the clinical data of the patients in the dataset opened for re-use in the current study.

Conflict of interest

The authors declared no conflict of interests.

Authors' contributions

Cui Z, Wang H, and Li CP designed the work. Wang C, Chang QX, and Wang XM analyzed the data, drafted and revised the article; Wang KY analyzed the data and drafted the article. All authors confirmed the final version of the work.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021; 71(3): 209-49. doi:10.3322/caac.21660.
2. Lee SE, Chung JS, Han BK, et al. Relationship of prostate-specific antigen and prostate volume in Korean men with biopsy-proven benign prostatic hyperplasia. *Urology* 2008; 71(3): 395-8. doi:10.1016/j.urol.2007.10.019.
3. Mousavi SM. Toward prostate cancer early detection in Iran. *Asian Pac J Cancer Prev* 2009; 10(3): 413-8.
4. The General Hospital of the People's Liberation Army. Prostate Cancer Data Set. Population Health Data Archive PHDA, 2019. CSTR: A0006.11.A0005.201905.000531.
5. Liu X, Cheng MH, Shi CG, et al. Variability of glomerular filtration rate estimation equations in elderly Chinese patients with chronic kidney disease. *Clin Interv Aging* 2012; 7: 409-15. doi:10.2147/CIA.S36152.
6. Taft LM, Evans RS, Shyu CR, et al. Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. *J Biomed Inform* 2009; 42(2): 356-64. doi:10.1016/j.jbi.2008.09.001
7. Toth R, Schiffmann H, Hube-Magg C, et al. Random forest-based modelling to detect biomarkers for prostate cancer progression. *Clin Epigenetics* 2019; 11(1): 148. doi:10.1186/s13148-019-0736-8.
8. Lan L, Wang Z, Zhe SD, et al. Scaling up kernel SVM on limited resources: A low-rank linearization approach. *IEEE Trans Med Imaging* 2019; 30(2): 369-78. doi:10.1109/TNNLS.2018.2838140.
9. Shin HC, Roth HR, Gao MC, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imaging* 2016; 35(5): 1285-98. doi:10.1109/TMI.2016.2528162.
10. Xie J. Study on the drying model of larch wood based on artificial neural network[dissertation]. Northeast Forestry Univer-

- sity; 2013.
11. The Online Platform for Prostate Cancer Risk Prediction. Department of Epidemiology and Health Statistics, Tianjin Medical University; c2021. Available from: <https://pcarisk.shinyapps.io/pcapred/>. Released: November 30, 2021.
 12. Lacher DA, Hughes JP. Total, free, and complexed prostate-specific antigen levels among US men, 2007-2010. *Clin Chim Acta* 2015; 448: 220-7. doi:10.1016/j.cca.2015.06.009.
 13. Ju HX, Wang T, Wang W, et al. A comparative study of prostate cancer prediction models. *Data Knowl Discov* 2021; 5(09): 107-14. doi:10.11925/infotech.2096-3467.2020.1185.
 14. Wang YF, Wu H, Xue WG, et al. Classification prediction and analysis of cancer risk factors for prostate cancer and prostate hyperplasia. *Acad J PLA Med Sch* 2021; 42(3): 277-81, +305. Chinese. doi:10.3969/j.issn.2095-5227.2021.03.008.
 15. Van Hemelrijck M, Garmo H, Holmberg L, et al. Prostate cancer risk in the Swedish AMORIS study: the interplay among triglycerides, total cholesterol, and glucose. *Cancer* 2011; 117(10): 2086-95. doi:10.1002/cncr.25758.
 16. Arthur R, Møller H, Garmo H, et al. Association between baseline serum glucose, triglycerides and total cholesterol, and prostate cancer risk categories. *Cancer Med* 2016; 5(6): 1307-18. doi:10.1002/cam4.665.
 17. Srikrishna G. S100A8 and S100A9: new insights into their roles in malignancy. *J Innate Immun* 2012; 4(1): 31-40. doi:10.1159/000330095.

(Edited by Liangjun Gu)

基于机器学习算法的前列腺癌风险预测和在线计算研究

王淳¹，常琴雪¹，王肖萌¹，王柯云¹，王鹤²，崔壮^{1*}，李长平^{1*}

¹ 天津医科大学公共卫生学院，天津 300070，中国

² 北京大学第一医院医学影像科，北京 100034，中国

摘要

目的 基于临床常用指标，采用机器学习方法构建前列腺癌风险预测模型，为前列腺癌的早期诊疗提供科学依据，评价人工智能技术在医疗健康数据平台下的应用价值。

方法 对国家临床医学科学数据中心提供的前列腺肿瘤预警数据集预处理后，使用平滑剪切绝对偏差 (smoothly clipped absolute deviation, SCAD) 算法筛选特征指标。采用随机森林 (Random forest, RF)、支持向量机 (support vector machine, SVM)、反向传播 (back propagation, BP) 神经网络、卷积神经网络 (convolutional neural network, CNN) 4 种模型预测前列腺癌发生风险，其中神经网络模型使用经 SMOTE 增强后数据拟合。不同模型的预测能力采用受试者操作特性 (ROC) 曲线下面积 (area under the curve, AUC) 进行比较。在确定最优模型后，使用 Shiny 开发前列腺癌风险预测在线平台。

结果 在预测变量中，除活检标本碎组织体积、血游离前列腺特异抗原 (fPSA) 外，无机磷、甘油三酯、游离钙等临床常用指标与前列腺癌也密切相关。在 4 种模型中，RF 预测效果最好 (准确率：96.80%；AUC：0.975，95%CI：0.964-0.986)，其次为 BP 神经网络 (准确率：85.36%；AUC：0.892，95%CI：0.849-0.934)，SVM (准确率：82.67%；AUC：0.824，95%CI：0.805-0.844) 与 BP 神经网络预测效果相近，CNN 预测能力最低 (准确率：72.37%；AUC：0.724，95%CI：0.670-0.779)。基于 RF 及预测指标成功开发了一种前列腺癌风险预测在线平台。

结论 本研究揭示了医疗信息化平台下传统机器学习方法和基础神经网络模型在疾病风险预测中的应用价值，为疑似前列腺癌并接受穿刺活检人群的前列腺癌预测提出了新思路。此外，开发在线预测系统有助于增强人工智能预测技术的实用性，使医疗应用更为便捷。

关键词：前列腺癌；随机森林；支持向量机；反向传播神经网络；卷积神经网络

*** 通讯作者：**崔壮 zhuangcui417@126.com；李长平 changpingli417@126.com